Pattern Recognition Letters 65 (2015) 75–80

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A robust framework for tracking simultaneously rigid and non-rigid face using synthesized data☆

Ngoc-Trung Tran[a,*], Fakhreddine Ababsa[b], Maurice Charbit[a]

[a] LTCI, Telecom ParisTech, 37-39 Rue Dareau, Paris 75014, France
[b] IBISC, University of Evry, 40, Rue du Pelvoux, Evry 91020, France

ARTICLE INFO

Article history:
Received 27 November 2013
Available online 9 July 2015

Keywords:
3D head tracking
3D face tracking
Rigid tracking
Non-rigid tracking
Synthesized face
Face matching

ABSTRACT

This paper presents a robust framework for simultaneously tracking rigid pose and non-rigid animation of a single face with a monocular camera. Our proposed method consists of two phases: training and tracking. In the training phase, using automatically detected landmarks and the three-dimensional face model Candide-3, we built a cohort of synthetic face examples with a large range of the three axial rotations. The representation of a face's appearance is a set of local patches of landmarks that are characterized by Scale Invariant Feature Transform (SIFT) descriptors. In the tracking phase, we propose an original approach combining geometric and appearance models. The purpose of the geometric model is to provide a SIFT baseline matching between the current frame and an adaptive set of keyframes for rigid parameter estimation. The appearance model uses nearest synthetic examples of the training set to re-estimate rigid and non-rigid parameters. We found a tracking capability up to 90° of vertical axial rotation, and our method is robust even in the presence of fast movements, illumination changes and tracking losses. Numerical results on the rigid and non-rigid parameter sets are reported using several annotated public databases. Compared to other published algorithms, our method provides an excellent compromise between rigid and non-rigid parameter accuracies. The approach has some potential, providing good pose estimation (average error less than 4° on the Boston University Face Tracking dataset) and landmark tracking precision (6.3 pixel error compared to 6.8 of one of state-of-the-art methods on Talking Face video).

1. Introduction

Three-dimensional (3D) face pose tracking in a video sequence involves the estimation of six rigid face parameters; the 3D translation and the three axial rotations.[1] This is an important issue, which has been receiving considerable attention [9]. 3D face tracking is useful in many domains, such as: video monitoring, human computer interaction and biometrics. The problem becomes more challenging if non-rigid face parameters, animation and/or expression, are also estimated, or when including illumination changes, the presence of many people, or occlusions. In this study, we adopt a classical approach, based on a set of local landmarks with the constraint of a 3D face model to track both rigid and non-rigid parameters of a single face from a monocular camera.

Since the pioneer work of [10,11], it is well-known that the Active Shape Model (ASM) and the Active Appearance Model (AAM) provide an efficient way to represent and track frontal faces. Many developments [16,23,39] improved the tracking in terms of fitting accuracy or profile view tracking. The Constrained Local Model (CLM) has been recently proposed [12] that consists of an exhaustive local search around landmarks constrained by a 3D shape model. Saragih et al. [29,36] improved this method in terms of accuracy and speed. Saragih et al. [29] were able to track a single face with vertical rotation of 90° in a well-controlled environment. Even though this face model is very efficient for face tracking, significant annotated data of pose directions are required to learn 3D model and appearance distributions. This is costly in unconstrained environments.

Several other face models have been considered, such as: cylinder in [7,25,40], ellipsoid in [3], and mesh in [33]. Largely, these methods could estimate the three rotations and even profile-view, but they only address rigid rather than non-rigid faces, making it impossible to accommodate facial expressions.

The popular 3D Candide-3 model has been defined to manage both shape and animation parameters. Ström [30] used the Kalman filter to target points of interest in a video sequence based on an

☆ This paper has been recommended for acceptance by J. Yang.

* Corresponding author. Tel.: +33 6 51 38 78 76.
E-mail address: tntrung@gmail.com, trung-ngoc.tran@telecom-paristech.fr (N.-T. Tran).

[1] As commonly used in the literature, we adopt the terms Yaw (or Pan), Pitch (or Tilt) and Roll for the three axial rotations.

http://dx.doi.org/10.1016/j.patrec.2015.06.028
0167-8655/

adaptive rendered keyframe. This approach is semi-automatic and insufficient for quick movement. Chen and Davoine [8] took advantage of local features constrained by this 3D model to capture both rigid and non-rigid head motions. However, this method worked poorly for profile-view due to the inefficiency of accommodating the large variability of landmarks. Ybanez-Zepeda et al. [42] proposed a linear model between the facial parameters and the appearance of face images. This method was only robust for face and landmark tracking on near frontal faces. Lefevre and Odobez [20] extended Candide face to work with profiles. Nevertheless, their objective function, combining structure and appearance features with dynamic modeling, appears to slowly converge due to the high dimension. Tran et al. [31] proposed an adaptive Bayesian approach to track principal components of landmark appearance. Their algorithm appears to be robust for tracking landmarks, but is unable to recover when tracking is lost. All these methods used a synthetic databases to learn training models.

A tracking framework is robust if it can work with a wide range of rotations, facial expressions, environmental changes and occlusions, and is also recoverable if tracking is lost. Cascia et al. [7,40] utilized dynamic templates of a cylinder model to address lighting and self-occlusion problems. However, this approach accumulated errors when accessing in long video sequences. Saragih et al. [29] considered the local features that are not much affected by the whole facial appearance, expressions and self-occlusion. For recovery, tracking-by-detection or wide baseline matching [17,33,35] was used to match the current frame with preceding keyframes. The matching is sufficient against fast movements and illumination changes, and is able to recover lost tracking. However, the matching is only suitable with rigid parameters. Moreover, these methods degraded when the number of keypoints detected on the face is too few. Recently, the cascaded regression has shown very impressive results in face alignment, such as [6,18,27,41], however, these method are developed for near frontal face between $\pm$ 45° of Yaw rotation. Asteriadis et al. [4] proposed the combination of traditional tracking techniques and deep learning to provide a proficient pose tracking. Many commercial products also exist, e.g. [14], which showed good results in pose and face animation tracking. But this product requires the controlled environment of illumination and movements. In addition, it must wait for the frontal view to recover when tracking is lost.

We propose a robust framework for tracking facial pose with large rotations, and facial animation. We focus on three aspects: (i) a synthetic database, (ii) SIFT descriptors and (iii) the combination of geometric and appearance models.

First, we build a synthetic database to avoid expensive and time-consuming manual annotation. This consists of a large set of face poses, approximately 6500 synthesized poses, with *Yaw*, *Pitch* and *Roll* from $-90°$ to 90° with 10° steps. This database provides a basis for non-rigid face and profile tracking.

Second, we utilize SIFT to represent the local patches. This is a well-known descriptor proposed by [22] discriminative and robust to illumination changes and to some affine transformations.

Third, a two-step approach is instigated: we first perform two-dimensional (2D) SIFT matching between the current frame and some preceding-stored keyframes to estimate rigid parameters only. Then, we obtain the whole set of parameters (rigid and animation) using maximum likelihood. A classical Huber M-estimator function was employed to robustify this approach. Our proposed system provides a very good compromise compared to other published algorithms on several public datasets, in terms of pose estimation and landmark localization.

The remaining of this paper: Section 2 describes the face model and the used descriptors. Section 3 discusses the pipeline of the proposed framework. Experimental results and analysis are presented in Section 4. Finally, in Section 5 we draw conclusions and discuss further perspectives.

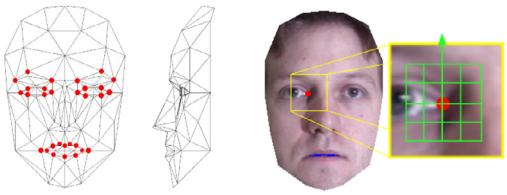

Fig. 1. Candide-3 frontal and profile view. Landmarks are used in tracking and defining the region to extract the SIFT descriptor of one landmark with fixed scale and direction.

2. Face representation

2.1. Shape representation

Candide-3, initially proposed by [2], is a very commonly used 3D model representing both facial shape and animation. It consists of $N = 113$ vertices representing 168 surfaces. If $g \in R^{3N}$ denotes the vector of dimension $3N$, obtained by concatenating of three components of the N vertices, the face model can be written:

$$g(\sigma, \alpha) = Rs(\bar{g} + S\sigma + A\alpha) + t \quad (1)$$

where $\bar{g}$ is the mean value of g. R is a 3D rotation matrix, s is a scale factor, and t is the 3D translation vector. The known matrices $S \in R^{3N\times 14}$ and $A \in R^{3N\times 65}$ are Shape and Animation Units that control shape and animation, respectively, through the σ and α parameters. Among the 65 components of animation control α, 11 track eyebrows, eyes and lips. Therefore, the full model parameter, b, has 17 dimensions: 3 of rotation (r_x, r_y, r_z), 3 of translation (t_x, t_y, t_z) and 11 of animation r_a:

$$b = [r_x\ r_y\ r_z\ t_x\ t_y\ t_z\ r_a]^T \quad (2)$$

Although σ and b are both calculated at the first frame, only b is re-calculated at the next frames because we assume that the shape parameters are unchanged later. During tracking, b at time t is re-written as b_t (Section 3.2.3).

2.2. Projection

We assume a weak perspective projection from three dimensions to two dimensions. Aggarwal et al. [1] showed that the focal length does not need to be accurately known if the depth between the 3D object and camera is much larger than the 3D object sizes. Therefore, the camera calibration has been obtained from empirical experiments. In our case, the intrinsic camera matrix is $K = [f_x, 0, c_x; 0, f_y, c_y; 0, 0, 1]$, where the focal length of camera is $f_x = f_y = 1000$ pixels and the coordinates of the camera's principal point (c_x, c_y) is the center of 2D video frame. Because of the perspective projection assumption, the depth, t_z, is directly related to scale parameter s.

2.3. Local representation

In our framework, a facial appearance is represented by a set of 30 landmarks (Fig. 1). The local patch of one landmark is described by SIFT descriptor. Because landmark positions are known, SIFT detector is unnecessary and only the SIFT descriptor is involved. Two important parameters need to be determined: SIFT scale and orientation when using toolbox [34] for SIFT descriptors. In this work, we define the scale of 1.2, which corresponds to a patch of approximately 15 $\times$ 15 pixels, the same size like some previous works [29,36], and the direction to be vertical. Note that face region is always normalized to 250 $\times$ 250 pixels before extracting the local descriptors, so there is no significant impact of depth translation.